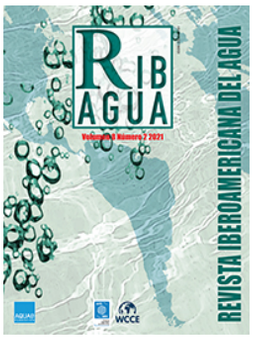

Ribagua
Revista Iberoamericana del Agua

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/trib20

El valor de los metadatos para las estaciones de recuperación de recursos del agua

Daniel Aguado, Frank Blumensaat, Juan Antonio Baeza, Kris Villez, María Victoria Ruano, Oscar Samuelsson, Queralt Plana & Janelcy Alferes

To cite this article: Daniel Aguado, Frank Blumensaat, Juan Antonio Baeza, Kris Villez, María Victoria Ruano, Oscar Samuelsson, Queralt Plana & Janelcy Alferes (2023): El valor de los metadatos para las estaciones de recuperación de recursos del agua, Ribagua, DOI: 10.1080/23863781.2023.2184286

To link to this article: https://doi.org/10.1080/23863781.2023.2184286

Published online: 06 Mar 2023.

RIBAGUA
https://doi.org/10.1080/23863781.2023.2184286

El valor de los metadatos para las estaciones de recuperación de recursos del agua

Daniel Aguado[a], Frank Blumensaat[b], Juan Antonio Baeza[c], Kris Villez[d], María Victoria Ruano[e], Oscar Samuelsson[f], Queralt Plana[g] and Janelcy Alferes[h]

[a]Institut Universitari d'Investigació d'Enginyeria de l'Aigua i Medi Ambient (IIAMA), Universitat Politècnica de València, València, Spain; [b]Institute of Environmental Engineering, Chair of Urban Water Management Systems, Zurich, Switzerland; [c]Department of Chemical, Biological and Environmental Engineering, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain; [d]Oak Ridge National Laboratory, Oak Ridge, TN, USA; [e]Chemical Engineering Department, Universitat de València, Burjassot, Spain; [f]IVL Swedish Environmental Research Institute, Stockholm, Sweden; [g]CentrEau, the Québec Water Research Center, Université Laval, 1065, Avenue de la Médecine, Québec, Canada; [h]VITO - Vision on technology, Belgium

RESUMEN

Los metadatos hacen referencia a información descriptiva (como ubicación del sensor, unidad de medida, rango de medida, fecha de calibración, fecha de limpieza, si ocurrió algún evento como episodio de lluvia/fallo operativo/vertido tóxico …) que es esencial para convertir los grandes volúmenes de datos que se recogen actualmente en las instalaciones de tratamiento de agua y que están sin procesar en información y recursos útiles. Con el avance de la digitalización en el sector del agua, es fundamental evitar los cementerios de datos y, por otro lado, utilizar los datos almacenados para resolver problemas actuales y futuros. Este artículo se centra en el papel crucial que tienen los metadatos para responder a desafíos futuros y posiblemente impredecibles. El objetivo de este documento es presentar el 'reto de los metadatos' y destacar la necesidad de tener en cuenta los metadatos cuando se recoge información como parte de las buenas prácticas de digitalización.

ABSTRACT

Metadata refers to descriptive information (such as sensor location, measurement unit, measurement range, calibration date, cleaning date, if any event occurred such as rain event/operating failure/toxic spill, …) essential to convert large volumes of raw data that are currently collected at water treatment facilities into useful information and resources. With the advance of digitalization in the water sector, it is fundamental to avoid data graveyards and, on the other hand, using collected data to address current and future problems. This paper focuses on the crucial role that metadata has in responding to future and possibly unpredictable challenges. The aim of this document is to present the 'metadata challenge' and to highlight the need to consider metadata when collecting information as part of good digitalization practices.

ARTICLE HISTORY
Received 9 July 2022
Revised 16 February 2023
Accepted 20 February 2023

KEYWORDS
Metadatos; digitalización; inteligencia artificial; aguas residuales; estaciones de recuperación de recursos del agua

1. Introducción

Para proteger de las aguas residuales tanto a los ecosistemas acuáticos naturales como la salud humana, las estaciones depuradoras de aguas residuales (EDAR) se han diseñado y operado tradicionalmente para eliminar los contaminantes que transportan estas aguas antes de su vertido final [1]. La transición hacia la Economía Circular ha supuesto importantes cambios a nivel mundial durante las dos últimas décadas, y en el caso de las aguas residuales ha motivado un cambio en la percepción de éstas que han pasado a considerarse un recurso [2] que permite la recuperación de agua limpia, nutrientes y energía. Con el objetivo de recuperar estos recursos valiosos del agua residual, muchas EDAR se han actualizado y modernizado mediante la incorporación e implementación de tecnologías de recuperación, de procesos innovadores y sostenibles, convirtiéndose de esta manera en estaciones de recuperación de recursos del agua (ERRA).

A medida que las EDAR y las ERRA entran en la era del *Big Data*, se enfrentan naturalmente a los desafíos de integrar actuadores inteligentes, sensores y sistemas de control autónomos de una manera sensata y transparente. Un aspecto que sigue siendo una carga importante para las empresas del sector del agua es el almacenamiento y la gestión de los datos procedentes de los sensores con vistas a su uso posterior. Para hacer posible una interpretación de los datos más allá del momento original de la recogida de los mismos, es fundamental que los datos almacenados de los sensores se

CONTACT Daniel Aguado daaggar@hma.upv.es Institut Universitari d'Investigació d'Enginyeria de l'Aigua i Medi Ambient (IIAMA), Universitat Politècnica de València, València, Spain

complementen con una descripción adecuada de los mismos (como ubicación del sensor, unidad de medida, rango de medida, fecha de calibración, fecha de limpieza, si ocurrió algún evento como episodio de lluvia/fallo operativo/vertido tóxico . . .), es decir, con metadatos. De hecho, los datos recogidos hoy pueden ser útiles en el futuro para responder a desafíos de operación aún más complejos y a nuevas demandas debidas a los impactos ambientales, la calidad del efluente producido y la eficiencia de los recursos. Dado que se desconocen estos desafíos futuros, es particularmente difícil definir los metadatos necesarios que permitan crear minas de datos útiles para el futuro, en contraposición a los actuales cementerios de datos, y garantizar su obtención y almacenamiento de manera oportuna. En este artículo, destacamos los aspectos más importantes de este reto de los metadatos y proporcionamos argumentos y soluciones tempranas que conducen a una recogida e interpretación armonizada de los datos. Este artículo representa el primero en español de muchos resultados del grupo de trabajo de la *International Water Association* (IWA) sobre la recogida y organización de metadatos (MetaCO Task Group), que cuenta con el apoyo de la IWA desde el año 2020. Este grupo de trabajo está elaborando actualmente un Informe CientíficoTécnico (Scientific and Technical Report, STR) sobre metadatos que complementará la información presentada en este artículo. Junto con este artículo, el STR (previsto para 2023) proporcionará información más extensa sobre el reto que supone los metadatos y la mejor manera de abordarlo.

Varios artículos de revisión [3–6]) sobre la aplicación de técnicas de análisis y procesado de datos existentes para extraer información valiosa de los datos recogidos en las EDAR y ERRA, coinciden en resaltar que a pesar de que existe una gran cantidad de métodos en la literatura y de que se han publicado muchos artículos, todavía su validación e implementación en instalaciones a escala real es muy limitada, incluso la falta de comprobación de la calidad de los datos que se recogen es frecuente en muchas instalaciones. Según se recoge en estos artículos de revisión, las técnicas de análisis de datos más utilizadas en EDAR son:

- Redes Neuronales Artificiales que se han utilizado en el contexto de las EDAR y ERRA para diferentes propósitos como la predicción del rendimiento del proceso, para el desarrollo de sensores inferenciales y para el desarrollo de algoritmos de control predictivo.
- Análisis de Componentes Principales, que se ha utilizado principalmente para la monitorización y detección de anomalías en el proceso, así como para mejorar la comprensión del proceso.
- Lógica Borrosa (fuzzy) que se ha aplicado para desarrollar algoritmos de control que se despliegan en muchas EDAR y ERRA.

En la actualidad, la gestión de datos en las instalaciones de tratamiento de agua está organizada de tal manera que los operadores de la instalación tienen a su alcance una cantidad abrumadora de datos, que es muy difícil de procesar y analizar de manera eficaz y oportuna para permitir una mejor comprensión y una toma de decisiones adecuada. Como consecuencia del esfuerzo que requiere analizar esta gran cantidad de datos, la información potencialmente valiosa que contienen permanece sin estar disponible y sin ser explotada. Además, el contexto de los datos recopilados (p. ej., las condiciones climáticas prevalecientes durante la toma de esos datos, si ocurrió algún problema con un sensor, cuándo se calibró el sensor, . . .) está mentalmente presente en la memoria de los operadores y técnicos del proceso sólo durante un período de tiempo limitado, y conforme más tiempo pasa sin haber analizado los datos menos información contextual está disponible para poderlos interpretar correctamente.

2. La necesidad de los metadatos

En las últimas décadas, el sector del agua ha experimentado una revolución en la instrumentación. Por ejemplo, la medición de la concentración de oxígeno disuelto se introdujo en las EDAR para aumentar las capacidades existentes de los sensores de flujo y de nivel en la década de 1980. Desde entonces, la variedad de sensores disponibles ha aumentado constantemente. Los desafíos y oportunidades en la recopilación del '*Big Data*' a menudo se clasifican en las siguientes cuatro 4Vs: Velocidad, Volumen, Variedad y Veracidad. Gracias a las técnicas de comunicación cada vez más eficientes y a las grandes reducciones en los costes de almacenamiento de datos, la recogida de datos se ha vuelto extremadamente escalable. Esto significa que las actuales EDAR y ERRA dominan las dos primeras de las cuatro Vs, la velocidad y el volumen [7].

Los desarrollos recientes en minería de datos, aprendizaje automático y optimización, potenciados por un poder computacional virtualmente infinito y algoritmos para el aprendizaje por ordenador, han sido recibidos con entusiasmo en el sector del agua. Muchos se sienten atraídos por las nuevas capacidades de toma de decisiones asistida por ordenador, tanto a nivel operativo como gerencial. Sin embargo, muchos intentos de avanzar en la automatización a partir de flujos de datos cada vez más grandes invitan a una dura confrontación con

las otras dos Vs del *Big Data*: variedad y veracidad [8, 9]. La inteligencia humana y las rutinas inteligentes siguen siendo necesarias para categorizar, estructurar, homogeneizar y convertir los datos en información valiosa. De hecho, este importante paso exige fácilmente el 40% de los costes en la mayoría de los proyectos de consultoría y ciencia de datos, tanto en el sector de tratamiento de aguas residuales como en otros sectores [10–12]. Este coste está asociado en gran medida a la necesidad de clasificar los datos disponibles (es decir, separar los datos adecuados y que sirven para el propósito que se persigue de los datos que no sirven) para evitar el problema común de *Garbage-In Garbage-Out*, que es actualmente más evidente que nunca.

Los autores de este artículo creen que el coste de esta tarea se podría reducir drásticamente si se actualizaran las prácticas rutinarias de recogida y gestión de datos para respaldar la toma de decisiones y la automatización. Más específicamente, los datos existentes deberían complementarse proporcionando información sobre el propósito original de su recogida, los dispositivos utilizados para su generación, la calidad de los datos y su contexto. Este tipo de información descriptiva se conoce como metadatos y es un ingrediente esencial para convertir grandes volúmenes de datos sin procesar en información procesable. De hecho, se necesita un conocimiento detallado de las mediciones realizadas para poder realizar un análisis de datos consistente y creativo, a fin de garantizar un impacto en las decisiones operativas y de diseño [12]. Desafortunadamente, en el sector del agua no existen pautas específicas disponibles para la producción, selección, priorización y gestión de metadatos.

3. Garbage-in garbage-out – ¿Por qué las herramientas de análisis de datos son exigentes?

Hay varias razones por las que simplemente recoger más y más datos no es suficiente. Esta afirmación es válida tanto cuando se utilizan modelos mecanicistas (por ejemplo, modelos basados en la física) como cuando se utilizan modelos empíricos (por ejemplo, los de aprendizaje automático y otros modelos basados en datos) para la interpretación de datos. A continuación, se muestran algunas de estas razones, así como las acciones necesarias para resolverlos:

(1) Los datos típicos incluyen mediciones de calidad sospechosa. Es común observar en las señales de los sensores los síntomas de errores de corta duración en forma de valores atípicos, picos, características de alto ruido y desviaciones de las mediciones de referencia que se mantienen en el tiempo. Además, los sensores utilizados en el sector del agua son propensos a fallos sistemáticos, a menudo debido a errores de calibración y deriva. Incluir todos los datos para crear un modelo de referencia sin eliminar o corregir las mediciones de baja calidad dará lugar a modelos defectuosos. Un requisito clave es que los datos utilizados para la creación e identificación del modelo sean de alta calidad, ya sea mediante la gestión adecuada del sistema de recogida de datos o mediante un proceso de refinamiento de los datos bien establecido, con herramientas de validación y conciliación de datos *on-line* u *off-line*.
(2) A menudo, los datos disponibles para la identificación del modelo no se corresponden con las condiciones en las que se implementará el modelo. Por ejemplo, los datos de caudal y composición del afluente de una ERRA en clima seco y en clima húmedo. Estos datos deben estar claramente separados para obtener un modelo representativo y fiable para el afluente típico de una ERRA bajo distintas condiciones de operación. Por tanto, otro requisito es que los datos utilizados para la identificación del modelo sean representativos de las condiciones de operación.
(3) Si bien se puede disponer de grandes volúmenes de datos, los patrones que uno pretende analizar son a menudo eventos poco frecuentes (por ejemplo, eventos ocasionados por la presencia de vertidos de tóxicos o por episodios de lluvia). De hecho, una de las muchas promesas de las herramientas de aprendizaje automático es que pueden ayudar a detectar y diagnosticar estos eventos. Para facilitar este proceso, debe estar disponible un volumen significativo de datos recopilados del patrón de comportamiento de interés o debe ser corregido cualquier desequilibrio en la frecuencia de los eventos de interés a través de la incorporación del conocimiento detallado del sistema.

Hoy en día, el requisito de recopilar datos de calidad conduce a una preselección necesaria pero laboriosa antes de poder aplicar el análisis de datos. Personas expertas analizan minuciosamente grandes volúmenes de datos y los modifican, los seleccionan y generan anotaciones para permitir una ejecución correcta de las tareas de optimización o predicción asistida por ordenador. Este es un esfuerzo tedioso y, a menudo, incluye evaluaciones subjetivas por parte de expertos

en el proceso que genera los datos. Los metadatos, si están disponibles, pueden ayudar. Primero, los metadatos informativos pueden ayudar en gran medida a automatizar de forma adecuada el triaje y selección de datos de calidad. En segundo lugar, los metadatos estructurados reducen la necesidad de una evaluación subjetiva, lo que a su vez aumenta la confianza en las predicciones algorítmicas y en la toma de decisiones. Dado que cualquier algoritmo basado en mediciones se basa en datos representativos, fiables e interpretables, los conjuntos de datos deben juzgarse en virtud de los metadatos proporcionados, junto con las medidas convencionales de la calidad de la señal del sensor, como veracidad, precisión y tiempo de respuesta [ver [13], para definiciones]. En la Figura 1 se ilustra cómo los valores de una medición por sí solos no son suficientes para aprovechar los beneficios de los sistemas intensivos de recogida de datos.

Con el propósito de realizar la selección y triaje de los datos, un buen conjunto de datos incluye metadatos sobre los siguientes aspectos:

(1) **Sistema de generación de los datos** – Descripción de la información de cada etapa del proceso de recogida de datos, incluyendo información sobre (a) el propósito de la recogida de los datos, (b) el hardware del sensor (por ejemplo, la medida y la resolución temporal, la unidad de medida, el principio de medida, el fabricante, el modelo del sensor, etc.), (c) el procesado de la señal incluyendo el registro, la transmisión, y el almacenamiento, y (d) el refinamiento de los datos incluyendo todas las transformaciones y modificaciones realizadas a los datos después de su recogida. Este tipo de metadatos permite la selección de los datos adecuados para el propósito y, a menudo, están disponibles desde el mismo dispositivo del sensor a través de un sistema de comunicación digital moderno (por ejemplo, Ethernet).

(2) **Calidad de los datos** – Información detallada sobre la calidad de la señal del sensor, basada en (a) el registro digital de los eventos de calibración, validación y verificación para cada sensor, (b) la descripción del registro individual de los datos que son sospechosos (por ejemplo, valores atípicos o picos) y (c) descripciones de periodos con datos de baja calidad (por ejemplo, debido a errores de calibración, falta de mantenimiento, desviaciones u otras disfunciones). Este tipo de descripciones pueden ser obtenidas a través de anotaciones manuales de un experto del proceso

Figure 1. ¿Podemos interpretar esta medida proporcionada igual a 42.0? Necesitamos saber mucho más para evaluar la información contenida en las señales de los sensores. El tipo de datos descriptivos que necesitamos para su interpretación se conoce como 'metadatos' e incluye el propósito de la medición, el principio de medida, sus unidades, el historial de mantenimiento realizado al sensor, los indicadores de calidad de la señal y el contexto espacial y temporal de la medida.

que genera los datos, pero también es importante un despliegue cuidadoso de las herramientas de análisis de datos, que a su vez se utilizan para la evaluación y el control de calidad de los datos cuantitativos. Esto significa que también deberían estar considerados e incluidos como metadatos (d) los datos producidos a partir de los métodos utilizados para recoger esta información (por ejemplo, análisis algorítmicos, protocolos de operación, anotaciones de expertos, etc.). Este tipo de metadatos permiten identificar datos que satisfacen la calidad requerida.

(3) **Información contextual** – Información que describe las circunstancias fuera de la instalación de tratamiento que pueden influir en la interpretación de las señales registradas en la misma. Puede tratarse de información sobre la forma de operación, la meteorología local, incluyendo cambios estacionales o las tormentas, o cambios significativos en la estructura y el modo de operación en las infraestructuras aguas arriba de la EDAR (por ejemplo, el alcantarillado). Esto también incluye información sobre eventos anormales y poco frecuentes, incluyendo fallos operativos, reuniones sociales o vertidos tóxicos. A menudo, este tipo de información es proporcionada por personal técnico u operadores. Además, este tipo de metadatos permite seleccionar datos que son relevantes para la tarea en cuestión, es decir, que es informativa y relevante.

Desafortunadamente, los tipos de metadatos descritos anteriormente raramente están disponibles. Por ejemplo, las descripciones de los procedimientos (como podría ser el mantenimiento de los sensores) puede faltar, los resultados de la validación del sensor pueden no estar registrados, y las circunstancias bajo las cuales los datos fueron recogidos (por ejemplo, durante una tormenta) pueden no ser conocidas. Este tipo de información es indispensable, sin embargo, para una predicción intensiva de datos y una optimización del rendimiento de las ERRA. Sin esta información, uno depende únicamente y de manera crítica de la memoria del personal a cargo de la instalación para interpretar los datos disponibles. En un año, los datos históricos pueden convertirse en inútiles para las tareas basadas en datos, ya que la memoria del personal se desvanece y los datos recogidos llegan a su fecha de caducidad. Esta pérdida de información descriptiva valiosa puede convertir las mediciones ricas en información en un cementerio de datos. Además, la falta de información descriptiva a menudo solo se hace evidente después de que ha transcurrido mucho tiempo desde que las medidas originales fueron recogidas. Por lo tanto, la gobernanza de datos no solo requiere que uno pueda responder las preguntas de hoy basándose en los datos, sino que también gestione los datos recogidos de forma que permitan responder de manera fiable a cuestiones futuras aún desconocidas. Para tener en cuenta estas incógnitas, así como el hecho de que la fuerza laboral envejece, para empoderar a los miembros del personal, para asegurar la utilidad a largo plazo de los datos históricos registrados (durante décadas) y para tomar decisiones basadas en los datos importantes a nivel operacional y de gestión, la recogida y el registro de los metadatos descritos anteriormente tendrían que convertirse en una rutina.

4. Estructuración de los meta-datos

Incluso cuando los metadatos recogidos se mantienen a salvo para su uso posterior, su utilización puede ser desafiante. De hecho, los metadatos residen frecuentemente en una amplia gama de especificaciones de diseño, manuales, protocolos y hojas de cálculo, a menudo almacenados y gestionados en distintas bases de datos y carpetas (como silos de información). La localización de los datos y su acceso a menudo es gestionada de manera ad hoc. Para permitir la clasificación de los datos de manera automática, los metadatos tienen que ser accesibles y estar almacenados de tal forma que permitan una navegación fácil. Para conseguirlo, los metadatos necesitan estar almacenados de manera estructurada y periódicamente actualizada. Cuando sea posible, será útil un sistema centralizado de almacenamiento de metadatos para gestionar el acceso a esta información, lo que asegurará su precisión e integridad.

La definición y la integración de los metadatos es a menudo un obstáculo y, por lo general, no forma parte de los productos de software disponibles en el mercado para el sector del agua. Por esta razón, es importante centrarse en las maneras en que esto se pueda conseguir. Esto incluye generar la identificación de una fuente primaria de datos, que consiste en la versión más completa de todos los datos y metadatos. Estos datos primarios actúan como fuente única de verdad y representan la comprensión compartida e inequívoca de los datos más actuales disponibles para todos los usuarios de dichos datos. Naturalmente, la identificación de la fuente primaria de datos implica una apreciación bien calibrada de la necesidad de tener una buena gobernanza de los datos, la información, los modelos, y el software. A su vez, esto significa que estos cambios afectan a casi todos los miembros de la organización/empresa, por lo que se requiere una alineación

cuidadosa de los objetivos y las necesidades como parte de las buenas prácticas de digitalización.

5. Mantenimiento de sensores – aprovechando las prácticas de evaluación y control de calidad existentes para obtener ganancias rápidas

Una gran cantidad de empresas de servicios públicos de aguas cuentan con protocolos para comprobar la validez de las señales de los sensores de manera rutinaria. En la mayoría de los casos, una medición de referencia es utilizada para determinar si se requiere una acción de mantenimiento, tal como una limpieza intensiva o una calibración. Sin embargo, este tipo de medición de referencia es raramente registrada.

Para ilustrar cómo una modesta mejora de las prácticas de evaluación y control de calidad existentes puede conducir a metadatos útiles, tomamos dos series de mediciones de la referencia [Ohmura et al. [14]]. En la Figura 2, se puede observar la compensación (*offset*) de dos sensores de pH en función del tiempo durante su operación durante 2 años. Esta compensación es el potencial medido del sensor en una solución de calibración con pH 7 y está disponible como parte de la curva de calibración del sensor almacenada dentro de un transductor típico. El sensor A muestra un perfil monótono y decreciente. Esta disminución se explica fácilmente por la deriva del electrodo de referencia. Esta desviación se compensa en la práctica mediante una calibración regular. Por el contrario, el sensor B presenta un aumento de la compensación después de 500 días. Esto se debe al desgaste irreversible del sensor y no se puede corregir mediante una calibración. Es importante destacar que ser capaces de distinguir entre la compensación realizada alrededor de 180 días de operación (−11,4 mV), que se explica por la deriva normal, y la compensación similar observada a los 510 días de operación (−11,5 mV), explicada como resultado del deterioro del sensor, solo es factible gracias al registro de los valores históricos de la compensación. Sin estos metadatos, tal diagnóstico no sería posible.

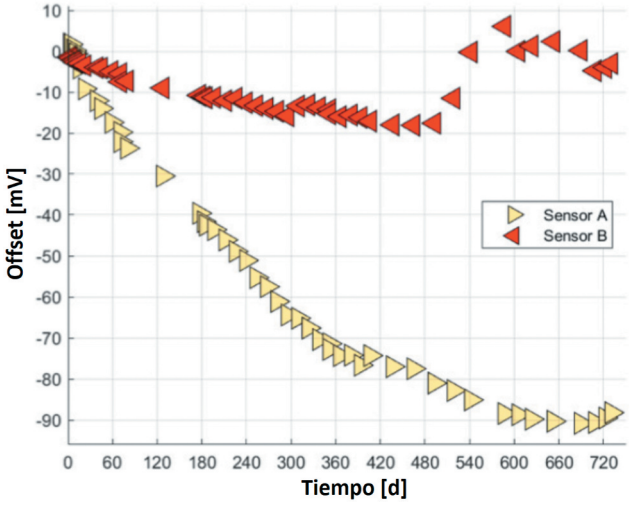

Figure 2. Efecto del desgaste en dos sensores de pH. Para el sensor A, la compensación (potencial del electrodo a pH = 7) disminuye gradualmente a medida que transcurre El Tiempo. Esta observación se atribuye a la deriva del electrodo de referencia y se puede corregir mediante la calibración. La deriva acumulada es de aproximadamente −90 mV al final de un período de dos años, lo que equivale a una variación de aproximadamente 1.5 unidades de pH en ausencia de calibración. En el sensor B, la compensación aumenta después de 500 días de uso. Esto se atribuye a daños irreversibles en el sensor y requiere el reemplazo del sensor. Este gráfico se ha generado con los datos tomados de [Ohmura et al. [14]].

Por lo tanto, recomendamos el registro sistemático de este tipo de metadatos antes y después de cada acción de mantenimiento ya ejecutada en la ERRA, incluida la limpieza, la calibración y el reemplazo de piezas. Esto permitirá una respuesta precisa y oportuna al desgaste del sensor, mejorando así la calidad de los datos. A largo plazo, también puede servir para disponer de información valiosa para implementar el mantenimiento preventivo del sensor, así como para decidir cuál es el mejor dispositivo de medición, particularmente al cuantificar el balance entre el coste del hardware del sensor versus la calidad de los datos obtenidos y los costes de las acciones de mantenimiento asociadas.

Como se mencionó previamente, una digitalización eficaz requiere la adopción y uso de buenas prácticas de gestión de metadatos, muchas de las cuales pueden automatizarse mediante una selección cuidadosa y una gestión estructurada de los metadatos. Esto ha creado una gama de roles de trabajo novedosos en el sector del tratamiento de aguas, con nombres como administrador de datos, ingeniero de datos o director de datos, lo que destaca la necesidad de experiencia interna en el manejo de datos, la gestión de expectativas con respecto a la transformación digital y la traducción de conceptos computacionales poco claros en un lenguaje de sentido común. Estos expertos pueden resultar de gran ayuda para convertir bases de datos obsoletas en una fuente eficaz de información a nivel operacional y para la toma de decisiones. Además, estos expertos deben estar estrechamente integrados en la fuerza laboral existente, para facilitar la pronta adopción de la toma de decisiones basada en las evidencias y para asegurar la alineación de expectativas y objetivos en toda la organización. En el futuro, esperamos que equipos y personal especializado puedan abordar los siguientes temas de relevancia, que actualmente son en general manejados por diferentes expertos en la materia, cada uno con su propia terminología aislada:

(1) Agregar e integrar nuevos dispositivos en un sistema de control existente.
(2) Gestionar y optimizar los sistemas de recogida de datos en entornos de usos múltiples, por ejemplo, datos para control en tiempo real, para informes, para la construcción y validación de modelos y para planificar actualizaciones importantes.
(3) Administrar y optimizar la calidad de los datos de sensores con la ayuda de herramientas de validación de datos básicas y avanzadas.
(4) Incrementar el flujo de datos existentes con metadatos de forma altamente automatizada.
(5) Fusionar las necesidades de metadatos para las operaciones de las ERRA, los requisitos algorítmicos y las rutinas actuales de recopilación de datos de sensores.

6. Ejemplo ilustrativo

Con objeto de ilustrar los beneficios de disponer de metadatos, se utilizó un conjunto de datos de una gran EDAR para estudiar la relación entre la tasa de nitrificación y el caudal de aire suministrado. La Figura 3 muestra un diagrama de dispersión de los valores registrados de ambas variables de interés: la tasa de nitrificación y el caudal de aire suministrado al reactor biológico de la EDAR. Como se puede ver en esta figura, no parece haber correlación entre ambas variables.

La Figura 4 muestra la evolución temporal de la concentración de amonio afluente a la EDAR durante el período estudiado. Se puede ver en esta figura que la concentración de amonio en la entrada experimentó importantes variaciones, con concentraciones desde menos de 10 mg/L hasta más de 40 mg/L. Afortunadamente, los datos registrados de la concentración de amonio estaban almacenados junto con metadatos, ya que se habían anotado diferentes anomalías que se produjeron a lo largo de los casi tres meses de funcionamiento que están representados. Gracias a los metadatos, los datos se han podido interpretar en su contexto permitiendo una mejor comprensión de los mismos. Los datos de la concentración de amonio registrados en condiciones normales o habituales de funcionamiento (es decir, excluyendo los períodos con anomalías) se seleccionaron y se utilizaron para entrenar un algoritmo de aprendizaje automático (Gaussian Process Regression (GPR)) para modelar la relación entre ambas variables (la tasa de nitrificación y el caudal de aire suministrado) así como para detectar valores atípicos (outliers) que no se pudieron descubrir inicialmente viendo los datos tal y como estaban representados en la Figura 3.

La Figura 5 muestra las predicciones y los intervalos de confianza del modelo GPR, de la relación entre la tasa de nitrificación y el caudal de aire suministrado. Los datos de test fuera de los intervalos de confianza para la predicción correspondientes a ± 2 veces la desviación estándar (línea continua negra delgada] generaron una alarma (puntos rojos). En esta figura se puede ver tanto en los datos de entrenamiento (puntos negros) como en los datos de test (puntos blancos) entre 15 y 45 g/s de amonio nitrificado la relación con el caudal de aire suministrado fue aproximadamente lineal. Es evidente que la Figura 5 es mucho más informativa que la Figura 3, y el motivo es la información descriptiva adicional contenida en los metadatos que ha permitido sacar mucho más partido de los datos registrados. Además, el modelo GPR desarrollado se puede utilizar en adelante para comparar nuevas medidas que se registren de la concentración de amonio con las predicciones del

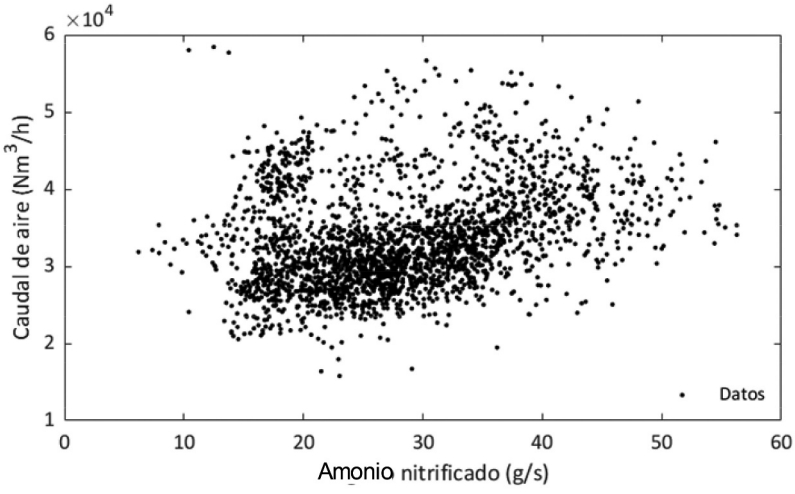

Figure 3. Diagrama de dispersión de la tasa de nitrificación y el caudal de aire suministrado al reactor aerobio de una EDAR. Figura adaptada de Samuelsson et al. [15].

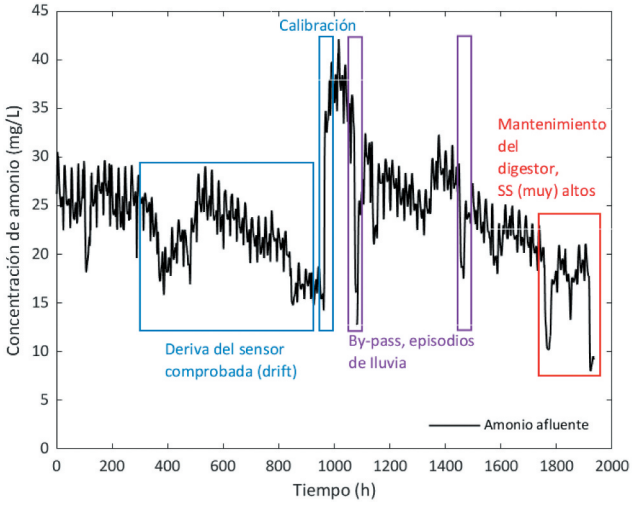

Figure 4. Concentración de amonio afluente a la EDAR con anomalías anotadas [metadatos). Resaltamos que esta figura sería mucho menos informativa sin las anotaciones mostradas. Figura adaptada de Samuelsson et al. [15].

modelo y de esta manera detectar desviaciones en la medición de amonio.

7. Conclusiones – *Take home messages* – Empezando con los metadatos

Tras las respuestas entusiastas a los avances en inteligencia artificial, robótica y aprendizaje automático, está cada vez más claro que cosechar los beneficios de una recogida intensiva de datos requiere añadir a las señales de los sensores información descriptiva. La necesidad de esta información descriptiva, denominada metadatos, y los desafíos que supone obtenerlos y gestionarlos refleja el hecho de que *there is no free lunch* (no hay almuerzo gratis). Una gobernanza de datos eficaz incluye la provisión de metadatos de alta calidad y marcará la diferencia entre fracasos y éxitos en sistemas basados en datos de monitorización, automatización y optimización. Como primer paso hacia gestión de datos, recomendamos a los administradores de las EDAR y de las ERRA lo siguiente:

1. Iniciar la automatización de la recogida de metadatos habilitando la integración de datos y el almacenamiento de metadatos de las acciones de mantenimiento del sensor (fecha de instalación, de calibración, de limpieza, de validación y de verificación).

2. Asegurar la disponibilidad de metadatos básicos para señales de sensores on-line en la misma ubicación que las señales de los sensores (por ejemplo, en la misma base de datos). Un conjunto muy básico de metadatos consta de:

a. Unidad de medida.
b. Rango de medición.
c. Resolución de la medida.
d. Principio de medición.
e. Ubicación del sensor.

3. Prepararse para prácticas avanzadas de metadatos, incluyendo la provisión de registros históricos completos de:

a. Los roles y/o propósitos del sensor.
b. Historial de valores medidos de compensación del sensor, de sensibilidad, de veracidad, de precisión y de tiempo de respuesta.

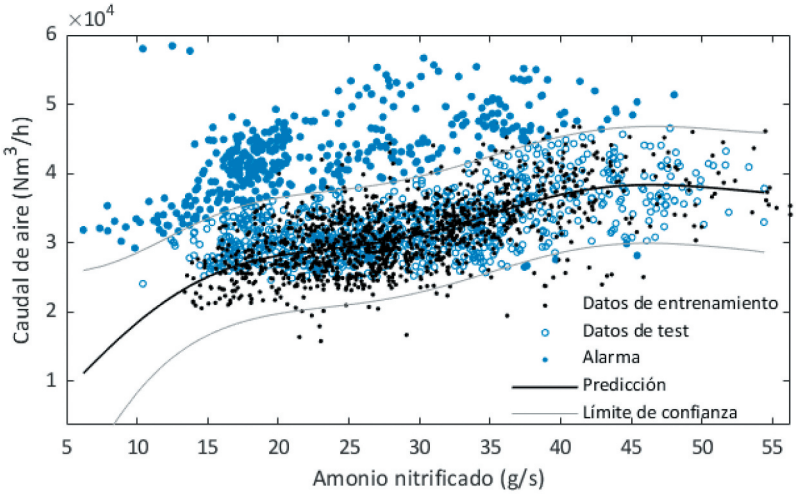

Figure 5. Diagrama de dispersión de la tasa de nitrificación y el caudal de aire suministrado al reactor aerobio de una EDAR, junto con la relación prevista a través del modelo GPR (línea continua negra gruesa) y su intervalo de confianza para la predicción obtenida como ± 2 veces la desviación estándar (línea continua negra delgada). Los datos de test que quedaron fuera del intervalo de confianza para la predicción generaron una alarma [puntos rojos). Figura adaptada de Samuelsson et al. [15].

c. Historial del estado operativo en que se encuentra el sensor (operativo, calibración, validación, mantenimiento).
d. Historial de protocolos de mantenimiento para la calibración y validación de sensores.

4. Evaluar el potencial de cualquier tipo de metadatos para evitar la caducidad de datos valiosos y convertir lo que actualmente son cementerios de datos en recursos valiosos para la toma de decisiones.

Agradecimientos

Los autores reconocen con gratitud las interesantes sugerencias de los profesores Vladan Babovic y Zoran Kapelan. Se agradece el apoyo tanto institucional como económico de la International Water Association (IWA) al grupo de trabajo de recogida y organización de metadatos (MetaCO Task Group).

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Metcalf & Eddy. Wastewater engineering: treatment and resource recovery, 5ed. 2013.978-0073401188
[2] Neczaj E, Grosser A. Circular economy in wastewater treatment plant–challenges and barriers. Proceedings. 2018;2(11):614.
[3] Corominas L, Garrido-Baserba M, Villez K, et al. Transforming data into knowledge for improved wastewater treatment operation: a critical review of techniques. Environ Model Softw. 2018;106:89–103.
[4] Newhart Kathryn B, Holloway Ryan W, Hering Amanda S, et al. Data-driven performance analyses of wastewater treatment plants: a review. Water Res. 2019;157:498–513.
[5] Jean-David T, Niels N, Vanrolleghem Peter A. A critical review of the data pipeline: how wastewater system operation flows from data to intelligence. Water Sci Technol. 2020 December 15;82(12):2613–2634.
[6] Phoebe M.L. Ching, Richard H.Y. So, Tobias Morck. Advances in soft sensors for wastewater treatment plants: a systematic review. Journal of Water Process Engineering. 2021;44. DOI:10.1016/j.jwpe.2021.102367.
[7] Sagiroglu S, Sinanc D. Big data: a review. 2013 IEEE International Conference on Collaboration Technologies and Systems (CTS), San Diego, 2013; 42–47. DOI:10.1109/CTS.2013.6567202.
[8] Ebbers M, Abdel-Gayed A, Budhi VB, et al. Addressing data volume. USA: Velocity, and Variety with IBM InfoSphere Streams V3. 0. IBM Redbooks; 2013.
[9] Normandeau K. Beyond volume, variety and velocity is the issue of big data veracity. Inside Big Data. 2013.
[10] Hauduc H, Gillot S, Rieger L, et al. Activated sludge modelling in practice: an international survey. Water Sci Technol. 2009;60(8):1943–1951.
[11] Kurgan LA, Musilek P. A survey of knowledge discovery and data mining process models. Knowl Eng Rev. 2006;21(1):1–24.
[12] Rieger L, Takács I, Villez K, et al. Data reconciliation for wastewater treatment plant simulation studies—planning for high-quality data and typical sources of errors. Water Environ Res. 2010;82(5):426–433.
[13] ISO. ISO15839: Water quality – On-line sensors/Analysing equipment for water – Specifications and performance tests. Geneva Switzerland: ISO; 2003.
[14] Ohmura K, Thürlimann CM, Kipf M, et al. Characterizing long-term wear and tear of ion-selective pH sensors. Water Sci Technol. 2019;80(3):541–550.
[15] Samuelsson O, Björk A, Zambrano J, et al. Gaussian process regression for monitoring and fault detection of wastewater treatment processes. Wat Sci Technol. 2017;75(12):2952–2963.